Mass Storage Elements of Data Intensive Computing as Exemplified by the High Performance Storage System (HPSS)

SC98 Tutorial on Data Intensive Computing

November 8, 1998

Orlando, Florida

The File Storage Group

National Energy Research Scientific Computing Center (NERSC)

Harvard Holmes, Keith Fitzgerald (leader), Jim Daveler, Wayne Hurlbert, James Lee, Nancy Meyer

Abstract

High performance archival storage systems typically must accommodate a variety of usage patterns while maintaining very high performance. We discuss typical usage profiles and their requirements, as well as unusual requirements to meet special needs. The HPSS (High Performance Storage System) provides Class-Of-Service (COS) mechanisms to customize the treatment of storage requests.

HPSS provides a distributed architecture which scales to the required performance levels. We review the architecture of HPSS. We review the expected evolution of mass storage devices and their likely performance levels. We describe some practical aspects of operating a Mass Storage System.

Outline

- The HPSS collaboration
- Requirements for Mass Storage Systems
- Architectures for Mass Storage Systems
- HPSS overall architecture
  - Network attached peripherals
  - Parallelism
  - Configuration flexibility
- Device Characteristics and Evolution for Mass Storage Systems
- Practical Aspects of Operating Mass Storage Systems

HPSS Organizational Structure

Executive Committee
(1 member per site)

Technical Committee
(1 member/developer)

Developers

User Sites

HPSS Collaborative Effort

HPSS is a successful collaboration including IBM, DOE Laboratories, and many users.

- Development Partners (and User Sites)
  - IBM Global Government Industry
  - Lawrence Berkeley National Laboratory (NERSC)
  - Lawrence Livermore National Laboratory
  - Los Alamos National Laboratory
  - Oak Ridge National Laboratory
  - Sandia National Laboratories
- User Sites (Partial List)
  - San Diego Super Computer Center
  - University of Washington
  - Fermi National Laboratory
  - NASA Langley Research Center
  - Commissariat a l'Energie Atomique (CEA)
  - Stanford Linear Accelerator Center
  - California Institute of Technology/JPL
  - Brookhaven National Laboratory
  - European Laboratory for Particle Physics (CERN)

HPSS Support

HPSS is a “service” from IBM, and not a “product.” This implies that HPSS does not go through all the release levels that “products” do -- it should reach the users’ sites more quickly

General Requirements for Mass Storage Systems

"The Role of the Super Computer is to Create I/O Bottlenecks."

- Absolute reliability of the users' data
- Speed, Speed, Speed!
- High Capacity
- Low Device Latency
- Low Network Latency
- Support for parallelism
- Support for commodity devices
- High availability, maintainability
- Self Healing-- automatic fault identification, isolation and bypass
- Support for Legacy Data
- Support for Technology Insertion
- Performance Monitoring and Tuning
- Account Management tools
- User Interfaces
- Operator Interfaces

Problematics of Super Computers

- Super Computers are commonly used to process very large jobs in a dedicated or semi-dedicated mode
- Jobs have phases in which the I/O demands reach very high peaks
- Multiple I/O channels must be used effectively
- The Storage System must interface to the highest performance disks, tapes and networks.

Performance Requirements for a Mass Storage System

- Network Speed: keep up with parallel disks (1 GB/sec or more)
- Disk Speed: utilize the full capabilities of the hardware (1 GB/sec or more)
- Tape Speed: utilize the full capabilities of the hardware (200MB/sec or more)
- Namespace Scalability to billions of objects (within a single system image/name space)

Operational Requirements of a Mass Storage System

Reliability and Stability

- Multiple copies of data: at least 4 copies allowed
- Metadata integrity and security: backup, logging, mirroring
- Broad user base: we'd like to see 10 sites with similar configurations to ours
- Operation in degraded mode: don't give up if you don't have to

Functionality and Extensibility

- APIs and libraries for a wide variety of clients

Management Provisions

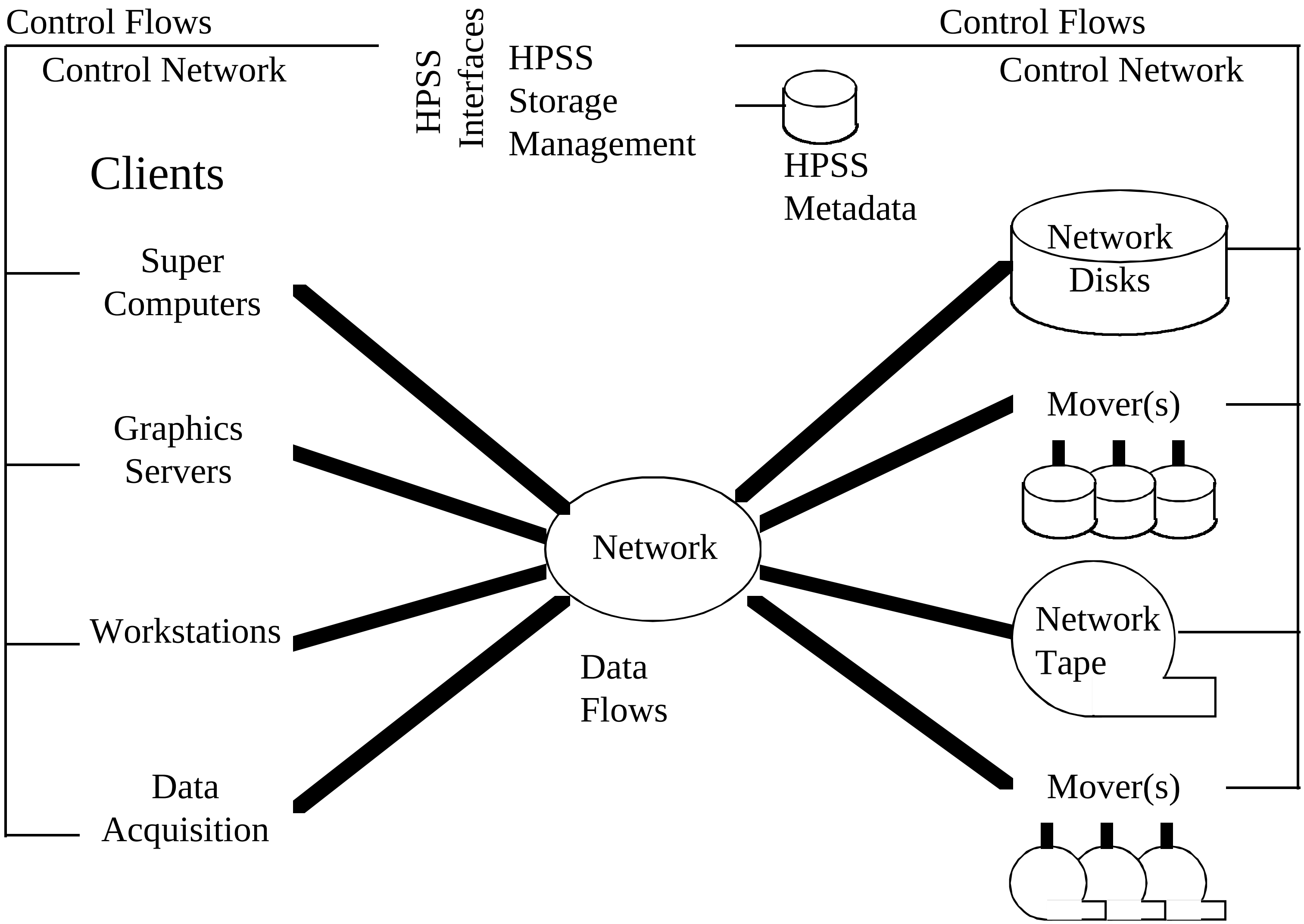
- Configuration ease: both initially and as the system changes and grows
- Monitoring ease: clear and specific status, operational and error messages

Architecture for Mass Storage Systems

THREE KEY IDEAS

- Network Attached Peripherals -- let data flow directly over the network between clients and storage devices, bypassing the hosts of the storage system
- Parallelism -- effectively use multiple, cost-effective machines (Redundant Array of Inexpensive Machines?).
- Flexible Configuration -- efficiently support a wide range of sizes, speeds and other attributes of stored objects

Overall Architecture of HPSS



Standards in HPSS

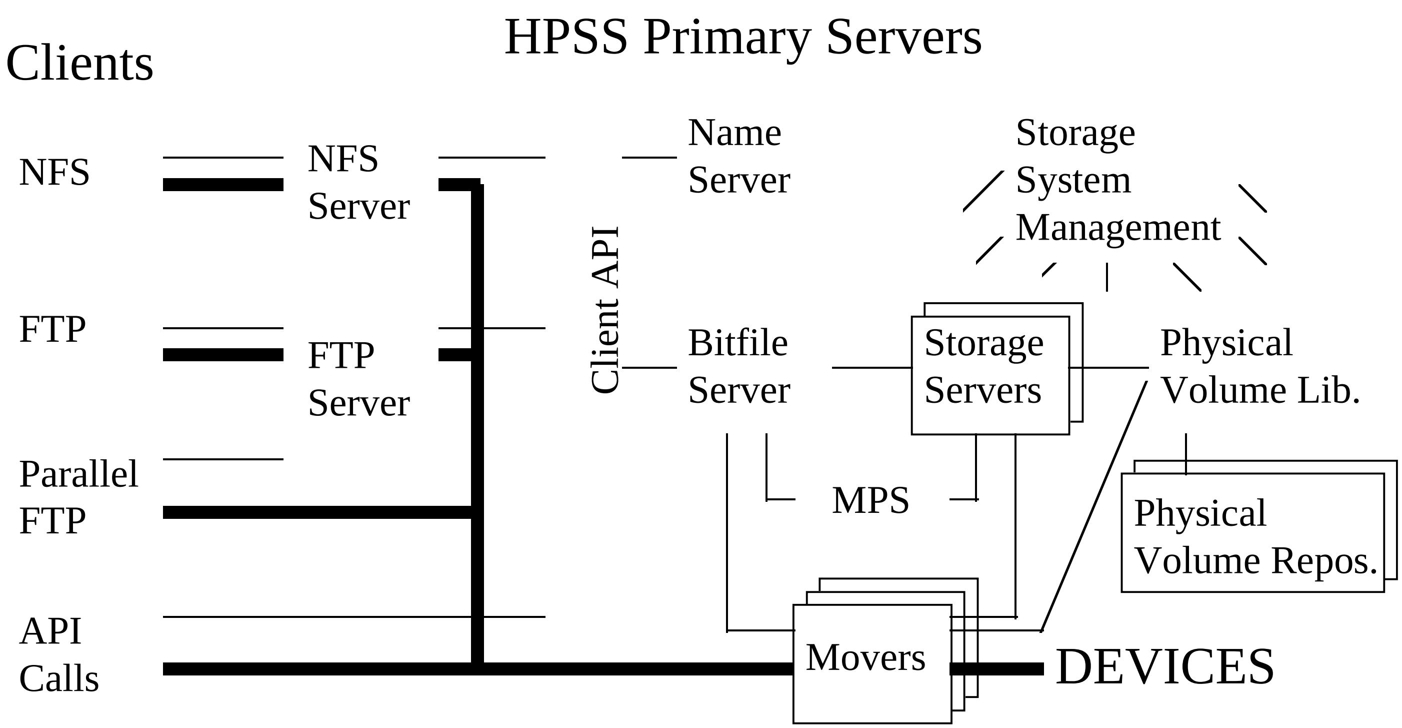
HPSS is designed around the extensive use of standards.

- The HPSS core architecture follows the IEEE Mass Storage Reference Model, Version 5
- HPSS implements Storage Server, Physical Volume Library, Physical Volume Repository, and Mover modules as defined in the IEEE Mass Storage Reference Model.
- HPSS uses standard commercially available products for its components where available.
  - Open Software Foundation Distributed Computing Environment for communication and security
  - Transarc's Encina product for transaction processing
  - Transarc's Encina Structured File System for metadata management.
  - Kinesiz's Sammi product for user interface display management

Overall Architecture of HPSS

- In theory, the storage management function is isolated from the high bandwidth data flows.
- The storage management component provides only control information, on a separate path, which can (and often is) a physically separate network.
- Control information is sent to the movers, which directly control the devices (i.e., disks and tapes). For network attached devices, control information is sent directly to the device (which is considered to have an integrated mover).
- The control information sets up the I/O transfer in the movers, thereby isolating the data on the device from any erroneous commands from the clients and maintaining the security of the stored data.
- After the I/O is set up, the client is instructed to start transferring data.

HPSS Primary Servers



<u>Infrastructure</u>
Logging and Accounting
GUI: Sammi
Metadata Management: Encina SFS
Transaction Management: Encina
Communication & Security: OSF DCE

HPSS Servers -- Components

- Name Server (NS): translates a human oriented name to an HPSS object identifier. It provides a standard POSIX view of the name space with directories, files, symbolic links and hard links.
- Bitfile Server (BFS): provides logical bitfiles to its clients. Bitfiles are identified by HPSS object identifiers. Clients may read and write portions of a bitfile by specifying the starting address and length. Reads and writes may be made in parallel to portions of a bitfile. A bitfile may have "holes" in its address space (where no data has been written). The BFS communicates with the Storage Server, which maps logical portions of bitfiles onto physical storage devices. The BFS supports migration, purging and staging of data in a storage hierarchy.
- Storage Server (SS): The SS provides several levels of storage objects: storage segments, virtual volumes and physical volumes. The SS maps references through the storage levels and handles the mapping of physical volumes into striped virtual volumes. The SS requests mounts and dismounts of physical media through the Physical Volume Library.

HPSS Servers -- Components (continued)

- Physical Volume Library (PVL): the PVL catalogues all HPSS physical volumes, mapping volumes to cartridges and cartridges to Physical Volume Repositories (media library managers). It allocates devices and cartridges, and performs atomic mounts of sets of requests to avoid deadlock. It issues commands to the Physical Volume Repositories to do physical mounts and dismounts.
- Physical Volume Repository (PVR): the PVRs manage all HPSS cartridges and interface with media libraries to perform mounts and dismounts. Multiple PVRs are supported and every cartridge must be managed by exactly one PVR.
- Mover (MVR): the MVR transfers data from a source device to a sink device, including optimizing requests and retries. Devices can have geometry, e.g., disk or tape, or not, e.g., network or memory. A single MVR can handle both ends of a data transfer, e.g., from disk to tape, if it can access both devices.

HPSS Servers -- Components (continued)

- Migration/Purge Server (MPS): the MPS manages the movement of data between storage levels in the storage hierarchies. Using the Bitfile Server and Storage Servers, the MPS copies data to lower levels in the storage hierarchy (data migration). The MPS makes duplicate copies of the data as it is copied, if requested. The purge component of MPS releases data from higher levels in the hierarchies after the data has been copied to a lower level. This is not done immediately, but only when additional free space is needed.
- Storage System Management (SSM): the SSM monitors and controls most aspects of HPSS operation, including configuration, startup and shutdown. It initiates one time operations such as starting system accounting, and it makes configuration changes during operation, such as adding or deleting cartridges or other system resources. The SSM receives and displays alarms and other events. It can retrieve and modify operating parameters and policies. It can display significant attributes in the HPSS tasks, such as space used, data rates, and others.

HPSS Infrastructure

- The HPSS infrastructure components provide the "glue" that holds HPSS together and they also provide some of the building blocks which support the server functionalities.
- Distributed Computing Environment (DCE): HPSS uses DCE from the Open Software Foundation as a basic infrastructure.
  - The DCE Remote Procedure Call mechanism is used for control messages,
  - the DCE Threads package is used for multitasking, and
  - the DCE Security and Cell Directory services are also used.
- Transaction management is provided by Encina from Transarc Corporation. Transaction management is used to support atomic operations which either complete or are backed out (rollback), as if they had not happened. This allows multiple operations to be synchronized to maintain a consistent state within HPSS. Encina also supports cleanup and recovery/abort of transactions when required.

HPSS Infrastructure (Continued)

- Metadata Management: the Encina Structured File System (SFS) is used to provide metadata storage and operations for HPSS metadata management. It provides rudimentary database functionality (B-trees, record and field level access, and primary and secondary keys). SFS is integrated with Encina's Transaction Monitor to provide data consistency.

- Logging: HPSS uses a centralized logging function which manages logging messages. As required, the Log Daemon may send log messages to the Storage System Management server.

- Accounting: A rudimentary accounting interface is provided. (In HPSS, much more work is needed here; the present system is mostly just a skeleton for "roll your own" snapshots. The next version, out about now, will maintain real time counters, incrementing and decrementing them based on storage system activity.)

HPSS User Interfaces

- Client API: this HPSS specific interface supports the full range of HPSS operations. In addition to the usual directory and file operations, it supports parallel I/O capabilities. It is the basis for all the other user interfaces.
- FTP: this is the industry standard FTP.
- Parallel FTP: the HPSS FTP server has extensions that allow data to be transferred in parallel across the network. A Parallel FTP (PFTP) client can establish multiple connections directly with multiple Movers in HPSS. On the client side the multiple connections can be to multiple clients. The PFTP client can specify the file offsets and lengths for each of the parallel transfers.
- Network File System (NFS) V2: this server provides standard NFS functionality to any industry standard NFS client. But performance of NFS isn't that great!
- Parallel I/O File System: HPSS acts as an external file system for use with the IBM SP2 PIOFS.

Steps in a Transfer Using Network Attached Peripherals

Example: Read from disk storage to a client

1. The Client gets the Object Identifier (OID) from the Name Server
2. The Client issues an open request to the Bitfile Server
3. The Client issues a read request (with an I/O Descriptor(IOD)) to the Bitfile Server
4. The Bitfile Server modifies the IOD to include storage segment information and issues a read request to the Storage Server.
5. The Storage Server translates the storage segment information to physical device addresses
6. The Storage Server issues a read device/write network request to the disk Mover which handles the disk. For example, for a Maximum Strategy disk array connected to HiPPI, this request is sent over an ethernet control network connection to the disk array. The read request includes the device addresses (this preserves security by not allowing the client to specify device locations to access).
7. The disk Mover creates a listen port and returns the address of the listen port to the Bitfile Server.

Steps in a Transfer Using Network Attached Peripherals (cont.)

8. The Bitfile Server builds a Write I/O Descriptor including the listen port information, then sends the write request and IOD to the client.

9. The client issues a read network request to its network I/O device driver, including the listen port of the disk Mover.

10. The Client network device driver connects to the disk Mover listen port and passes transport mechanism, length and offset.

11. The disk Mover creates a data port, if necessary, and returns the address of the data port.

12. The Client network driver sends a read request to the data port and transfers the data read to the requester's buffer.

Steps 9 through 12 are repeated as needed.

-----

13. The Client issues a close to the Bitfile Server

Steps in a Parallel Transfer

- A Parallel Transfer adds to the above process the creation of multiple I/O Descriptors which are sent to multiple Movers.
- The Client must arrange to partition up the I/O Descriptors among its I/O facilities and have each facility connect with the corresponding Mover(s).
- It is both possible and reasonable for the Client to be composed of multiple hosts, all transferring parts of the same file at the same time.
- Parallel transfers usually involve striped files, but it is not required. With striped files, it will generally be optimum for transfers if the striping configuration is the same on both sides of the transfer.
- The hard part of parallel transfers (not covered here) is the resource allocation, synchronization and locking required to avoid deadlock and achieve high performance.

Multiple Configurations to Optimally Handle Files with Disparate Attributes

Multiple configurations for multiple service requirements are managed using the following concepts:

Class of Service → Storage Hierarchy → Storage Class 1

Storage Hierarchy → Storage Class 2

Storage Hierarchy → Storage Class 3

Managing Multiple Configurations

- Storage Classes are used to define the logical and physical organization of each particular type of storage device. Parameters include Media Block Sizes, Virtual Volume Block Sizes, Stripe Width, Migration Policy, Purge Policy and numerous others.
- A Storage Hierarchy is a list of Storage Classes used to store files. Files always reside in a single Storage Hierarchy. A single file may have segments at different levels of the Storage Hierarchy. As files age they are generally moved (migrated) down the levels of the hierarchy by the Migration/Purge Server.
- A Class of Service (COS) provides a user view of a Storage Hierarchy, including hints about the performance, limitations and use of the particular class of service. Some utilities provide automatic selection of a COS, based on the file size to be stored.

Using Multiple Configurations

- A typical installation will have from 3 to 10 Classes of Service. Typical configurations would include:
- A COS for small files, up to 10 MB or so, using small block sizes, smaller, slower devices, and unstriped media, designed to be space efficient.
- A COS for medium sized files, 10 MB to 1 GB, using larger block sizes, designed for medium performance.
- A COS for large files, with large block sizes, faster devices, and striping, designed for maximum performance.
- A COS for backup operations. Backup datasets are best kept separate from others because their behavior is write mostly and they have a short lifetime (3 to 6 months).
- Other special purpose COSs as needed. Conversions from UniTree or other storage management systems are often put into their own class of service for ease of management (setting them read only, repacking the media, etc.).

DISK PERFORMANCE TRENDS

	Size	Sequential Speed
Year 2000	18 GB	30 MB/sec
Year 2002	40 GB	50 MB/sec
Year 2004	80 GB	75 MB/sec

Observations on Disk Performance Trends

- The trends will probably continue for at least the next 5 years.
- Device latency is improving, but at a rather slow rate.
- Disk controllers will get smarter and smarter, as their computational power increases. They will become increasingly network centric.
- Utilization of the computational power in disk drive controllers is progressing -- even performing data base operations in the controller and only passing wanted data to the host.

Tape Performance Trends

Year 2000	40 GB	15 MB/sec
Year 2002	80 GB	30 MB/sec
Year 2004	160 GB	60 MB/sec

For one perspective on tape performance trends, see

http://www.LTO-technology.com/

Observations on Tape Performance Trends

Tape devices seem to have a three stage life cycle:

- a new media/head/performance combination is designed
- the media length is increased,
- the track/bit density is increased

After these three stages, a new design with incompatible media is generally required. So you can look for these stages as you buy new tape drives.

Tape devices are adopting disk drive technology:

- track following servos
- increased track and bit density requiring higher coercivity media
- linear tapes are maintaining positioning information to allow "seeks" to much more quickly reach the desired position.

Because of this, and because the bit density of disk drives is much greater than tape drives, there is still a long growth path left for tape drives.

Tape manufacturers are looking for ways to improve the seek time to match improved data rates. Linear tapes seem to have some advantage in seek and load times compared to helical scan tapes.

Observations on Optical Tapes

They always seem to be "just around the corner."

Magnetic tape is improving at a rate that makes it a very hard target to beat.

As an example of vendor aspirations, see LOTS Technology, Inc. (http://www.laser.com). They are aiming for

- 1 TB per cartridge
- 25 MB/sec sustained data transfer rate
- '3840-size' cartridge compatible with existing robots
- Delivery in approximately two years

What about DVDs?

Too early to tell...

Practical Aspects of Operating a Mass Storage System

The following aspects are in no particular order…

- **Mass storage systems take about 10 years to mature. If your storage system is younger than this, expect a lot of problems.**
- **It's very difficult to make a distributed file system as fast as a local one. A major obstacle is increased latency for file metadata operations--creates, opens, closes and status checks. The semantics of these operations are used for synchronization by applications, so they can't be changed. For local file systems, the file metadata is cached in main memory, so operations take only a few microseconds. For distributed file systems, many of these operations require a network inquiry to preserve correct semantics, so they take orders of magnitude longer.**
- **Market forces have dictated that high performance systems must get their performance from lots of commodity devices. The trick to performance is managing the required parallelism.**

Practical Aspects of Operating a Mass Storage System (cont.)

- Change is constant. Your days will be filled with endless configuration changes. Make sure that your system will make these as easy as possible.
- You can never get rid of your users' data, or back it up and restore it. (Your system is the backup system!) This will increase the risk of all of your configuration changes. Any new system you contemplate must have a migration path to preserve your users' existing data. This includes any hardware/software changes to your existing system.
- Repair procedures must be simple. The operators have a million things to do, even in the middle of the night.
- The system must withstand erroneous repairs--people do the most obvious thing first, and it isn't always right. Anticipate common operator errors.
- Error messages that advise how to fix the problem are nice, but in complicated systems, most significant problems have several potential causes. Operators quickly learn to recognize the common problems.

Practical Aspects of Operating a Mass Storage System (cont.)

- Log files are critical to let the staff see what's happening internally. This is critical to trouble shooting, and invaluable to performance analysis. By the time your system is up and running, it is unlikely you can duplicate operational problems in a test environment. Your only hope is the log files.
- "Sleepers" (a NERSC term) are flags on all the major client machines which tell your clients that the system is temporarily down--wait rather than returning an error. Doing this on client machines means that it is effective even when the storage server is completely down. The ability to pause a client job rather than returning an error allows a lot of rework to be avoided.
- In a large installation, the key to managing the workload is to segment it according to its characteristics, and then optimize each segment individually. For example, our backups get recycled every 90 days, so we segment out backups and put them on their own tapes. Since these files are released after 90 days, the tapes never need to be repacked. In fact, since the backup files do not need to be accessible by the general user, we have put them on a separate system. This makes optimizing their configuration even easier. We also use classes of service to separate out small files and put them on smaller slower devices because transfer rate is not such an issue for small files compared to latency. There tends to be many, many small files and they need smaller allocation chunks than large files, if they are not to consume a disproportionate share of the storage space.

More Information

1. Mike Gleicher, "HPSS", "http://www.sdsc.edu/hpss/", 1997; perhaps the introduction to HPSS on the Web.

2. IBM, "High Performance Storage System," http://www5.clearlake.ibm.com:6001/; the focus of development efforts.

3. IEEE Project 1244, Reference Model for Open Storage Systems Interconnection, Lester Buck et al., eds., Sept. 1994.

4. HPSS System Administration Guide, October 1997,

5. Rajesh Agarwalla, Madhu Chetuparambil, Craig Everhart, T. N. Niranjan, Rena Hayes, Hilary Jones, Donna Mecozzi, Bart Parliman, Jean E. Pehkonen, Richard Ruef, Benny Wilbanks, Vicky White, "HPSS/DFS: Integration of a Distributed File System with a Mass Storage System," Sixth Goddard Conference on Mass Storage Systems and Technologies, University of Maryland Conference Center, College Park, MD, March 23-26, 1998, pg. 57

More Information

6. Gordon Knight, "Near Field Recording, TeraStor's Mass Storage Benchmark Technology for the Next Decade," Sixth Goddard Conference on Mass Storage Systems and Technologies, University of Maryland Conference Center, College Park, MD, March 23-26, 1998, pg. 125

7. Jamie Shiers, "Building a Database for the Large Hadron Colider (LHC): the Exabyte Challenge," Sixth Goddard Conference on Mass Storage Systems and Technologies, University of Maryland Conference Center, College Park, MD, March 23-26, 1998, pg. 385

8. LOTS Technology, Inc. http://www.lasertape.com, June 1998

9. LTO-Technology (Linear Tape-Open), Hewlett Packard, IBM, and Seagate, August 1998, http://www.LTO-technology.com/

More Information

10. National Media Laboratory, “Optical Tape Technology Final Report”, 1995; http://www.nml.org/Publications/TechnicalReports/TechnologyAssessments/OpticalTapeTechnology/index.html; somewhat dated information on Optical Tape Technology

11. Rob Mairs, "MSS Requirements (?)," 1995 Goddard MSS Conference

12. Open Software Foundation, "Avoiding Gridlock on the Information Highway," 1994; a white paper describing how DCE is used in the HPSS architecture, and mentioning the National Storage Lab.

13. Danny Teaff, Dick Watson, & Bob Coyne, "The Architecture of the High Performance Storage System," 4th NASA Goddard Conference on Mass Storage Systems and Technologies, University of Maryland Conference Center, College Park, MD, March 28-30, 1995; white paper, includes 22 references

More Information

14. Samuel S. Coleman, Richard W. Watson, Robert A. Coyne, Harry Hulen, "The Emerging Storage Management Paradigm," Twelfth IEEE Symposium on Mass Storage Systems, Monterey, CA, April, 1993, p.p. 101; towards a new paradigm of network based storage systems, with challenges in high-speed devices, communications, and parallel computing. 19 references

15. Robert Heyer, Rich Reuf, Richard Watson, "High-Performance Data Transfers Using Network-Attached Peripherals at the National Storage Laboratory," Twelfth IEEE Symposium on Mass Storage Systems, Monterey, CA, April, 1993, p.p. 275; white paper on NSL work on 3rd party IPI transfers over HIPPI, fairly technical. 10 references.

16. Richard W. Watson, Robert A, Coyne, "The Parallel I/O Architecture of the High Performance Storage System (HPSS)," Fourteenth IEEE Symposium on Mass Storage Systems, Monterey, CA, September, 1995, p.p. 27, URL: http://www.computer.org/conferen/mss95/watson/watson.htm; basic work on parallelism in HPSS, 48 references.

More Information

17. Dick Dixon, " Changing Horses Mid-stream (Or, How Do You Follow One of the Most Successful Acts in Mass Storage?)", Fourteenth IEEE Symposium on Mass Storage Systems, Monterey, CA, September, 1995, p.p. 2; a large effort to develop a requirements document for mass storage systems.

18. European Mass-Storage Specification Working Group, "Statement of Requirements," working version 1.2, October 3, 1994. ftp://ftp.ecmwf.int:/pub/mssspec/mssspec.v1.2. A 74 page compendium of detailed requirements. See summaries in ref. 7.

19. Infotech SA Inc., The Mass Storage Report, Silver Spring, MD, Spring 1996. An industry review of mass storage offerings of both hardware and software. 2 volumes, ~400 pages. http://www.access.digex.net/~infotech/;

20. Thomas S. Woodrow, "Hierarchical Storage Management System Evaluation," NASA Systems Development Branch, NASA Ames Research Center, Report RND-93-014, August 1993; http://www.nas.nasa.gov/NAS/TechReports/RNDreports/RND-93-014/RND-93-014.html. Review of NAStore, FileServ, DMF and UniTree.

More Information

21. Michael F. Shields, "Robotic Tape Library System Level Testing at NSA, Present and Planned," 4th NASA Goddard Conference on Mass Storage Systems and Technologies, University of Maryland Conference Center, College Park, MD, March 28-30, 1995; STK Silo's have worked well; Convex/EMASS with ER90 Helical Scan drives in an Odetics Data Tower worked poorly; Sun/AMASS/Metrum RS48 robotic tape system was stable for low performance; discovered that EMASS FileServ HSM has scalability limits due to Ingress.

22. C. Linett, S. Ranade, "The Real Problems of Backup", 4th NASA Goddard Conference on Mass Storage Systems and Technologies, University of Maryland Conference Center, College Park, MD, March 28-30, 1995; metadata is cited as the hardest problem. EMASS FileServ and UniTree from UniTree are cited as coming closest to a solution.

23. Robert A. Coyne, Harry Hulen, "An Introduction to the Mass Storage System Reference Model, Version 5", Twelfth IEEE Symposium on Mass Storage Systems, Monterey, CA, April, 1993, p.p. 47.

More Information

24. Kevin C. Matthews, "Implementing a Shared File System on a HIPPI Disk Array," Fourteenth IEEE Symposium on Mass Storage Systems, Monterey, CA, September, 1995, p.p. 77; Cray's Shared File System, how it works, and a good description of i-node locking issues and how to get good performance; interactions with DFS and DMF.

25. Peter R. Berard, Thomas L. Keller, "Scientific Data Management in the Environmental Molecular Sciences Laboratory," Fourteenth IEEE Symposium on Mass Storage Systems, Monterey, CA, September, 1995, p.p. 162; an OODBMS on top of a mass storage system, initially UniTree; this presentation predates PNNL's choice of FileStor for its HSM.

26. Donald Creig Humes, Juliet Z. Pao, "Experiences from NASA/Langley's DMSS Project," Fourteenth IEEE Symposium on Mass Storage Systems, Monterey, CA, September, 1995, p.p. 182; requirements; NSL UniTree was chosen, partly as less costly; performance reviewed; some deficiencies noted in file manipulation operations and in system management utilities; comparison with DMF – DMF is more mature and more stable, but UniTree is acceptable.

More Information

27. Betty Jo Armstead, Stephen Prahst, "Implementation of a Campus-wide Distributed Mass Storage Service: The Dream vs. Reality," Fourteenth IEEE Symposium on Mass Storage Systems, Monterey, CA, September, 1995, p.p. 190; for NASA Lewis Research Center - they chose UniTree because it was "open"; the usual lessons learned - UniTree bogs down on lots of small files.

28. Mark Geary, Barry Howard, Steve Louis, Kim Minuzzo, Mark Seager, "Cooperative High-Performance Storage in the Accelerated Strategic Computing Initiative," Fifth NASA Goddard Conference on Mass Storage Systems and Technologies, University of Maryland Conference Center, College Park, MD, September 17-19, 1996, pp. 21; the authors are from LLNL; seeks "to integrate multiple autonomous HSMs into a cohesive whole;" DFS is a key file system interface; a back end HSM will be integrated via the DMIG DMAPI specification; DFS will be extended for parallel transfers; nameservers will follow links to other nameservers to provide a single system name space.

29. Terry Jones, Richard Mark, Jeanne Martin, John May, Elsie Pierce, Linda Stanberry, "An MPI-IO Interface to HPSS," Fifth NASA Goddard Conference on Mass Storage Systems and Technologies, University of Maryland Conference Center, College Park, MD, September 17-19, 1996, pp. 21; at LLNL, uses 3rd party IO over Fibre Channel; describes distributed transfers.

More Information

30. Rita Chambers, Mark Davis, "Petabyte Class Storage at Jefferson Lab (CEBAF)," Fifth NASA Goddard Conference on Mass Storage Systems and Technologies, University of Maryland Conference Center, College Park, MD, September 17–19, 1996, pp. 77; the current implementation uses Open Storage Manager (OSM) Hierarchical Storage Management software (Computer Associates, International), with Fibre Channel RAID disks; has analysis of data flow requirements, drive choices (they use Storage Tek Redwood drives), and network choices.

31. Ellen Salmon, "Storage and Network Bandwidth Requirements Through the Year 2000 for NASA Center for Computational Sciences," Fifth NASA Goddard Conference on Mass Storage Systems and Technologies, University of Maryland Conference Center, College Park, MD, September 17–19, 1996, pp. 273; data storage needs scale with computer power; peak transfer loads can be 20x average loads; the TB added per year is about twice the GFLOPS computing capacity; good information on working sets, read/write ratios, etc.

Sources (continued)

32. Ethan L. Miller, "Towards Scalable Benchmarks for Mass Storage," Fifth NASA Goddard Conference on Mass Storage Systems and Technologies, University of Maryland Conference Center, College Park, MD, September 17-19, 1996, pp. 515; benchmarks must contain several tests, so one of them will approximate the workload of the potential user; short-running benchmarks stress the maximum performance of the system, long-running benchmarks measure file migration algorithms and other long running processes (long benchmarks will often be simulated); some possible benchmarks are suggested.

33. Theodore Johnson, "Queuing Models of Tertiary Storage," Fifth NASA Goddard Conference on Mass Storage Systems and Technologies, University of Maryland Conference Center, College Park, MD, September 17-19, 1996, pp. 529; a detailed model of Robotic Storage Libraries; gives insight into striping vs. clustering; striping improves transfer rate, but increases drive contention, clustering reduces drive contention - for heavy loads, lean toward clustering. A Web pointer was given, which included source code.

More Information

34. W. S. Oakley, "Progress Toward Demonstrating A High Performance Optical Tape Recording Technology," Fifth NASA Goddard Conference on Mass Storage Systems and Technologies, University of Maryland Conference Center, College Park, MD, September 17-19, 1996, pp. 571; from LOTS Technology; aiming at a 1 TB capacity per cartridge (3480 size).

35. Kent Angell, "SAM-FS -- LSC's New Solaris-Based Storage Management Product," Fifth NASA Goddard Conference on Mass Storage Systems and Technologies, University of Maryland Conference Center, College Park, MD, September 17-19, 1996, pp. 593; sales pitch for SAM-FS, a Solaris based file system: fast restore of i-nodes, open archive format (tar files), flexible migration and staging, good (native) tape and disk performance. In use at DLR in Germany (equivalent to NASA in USA).

36. HPCwire, "From Petabytes to Gigabytes," article 70055, November 20, 1996; a review of Sony's tape offerings: DIR-1000 Series 19mm drives (up to 64 MB/sec, 10^(-13) error rate), ADI-1150 half inch data recorder (up to 15 MB/sec, 51.8 GB/cassette uncompressed, 10^(-13) error rate), DTF half inch tape drive (12 MB/sec, 42 GB/cassette uncompressed, 10^(-17) error rate); the PetaSite tape library accepts all three types of drives, and scales up to 3,000 TB.